Transfer matrix solution of the Wako-Saitô-Muñoz-Eaton model augmented by arbitrary short range interactions

V I Tokar[1,2] **and H Dreyssé**[1]

[1]IPCMS, UdS–CNRS, UMR 7504, 23 rue du Loess, F-67034 Strasbourg, France
[2]Institute of Magnetism, National Academy of Sciences, 36-b Vernadsky Boulevard, 03142 Kiev-142, Ukraine

E-mail: tokar@ipcms.u-strasbg.fr

Abstract. The Wako-Saitô-Muñoz-Eaton (WSME) model, initially introduced in the theory of protein folding, has also been used in modeling the RNA folding and some epitaxial phenomena. The advantage of this model is that it admits exact solution in the general inhomogeneous case (Bruscolini and Pelizzola, 2002) which facilitates the study of realistic systems. However, a shortcoming of the model is that it accounts only for interactions within continuous stretches of native bonds or atomic chains while neglecting interstretch (interchain) interactions. But due to the biopolymer (atomic chain) flexibility, the monomers (atoms) separated by several non-native bonds along the sequence can become closely spaced. This produces their strong interaction. The inclusion of non-WSME interactions into the model makes the model more realistic and improves its performance. In this study we add arbitrary interactions of finite range and solve the new model by means of the transfer matrix technique. We can therefore exactly account for the interactions which in proteomics are classified as medium- and moderately long-range ones.

PACS numbers: 05.50.+q, 87.15.hm, 68.43.De

1. Introduction

The WSME model is a generalization of the one dimensional (1D) lattice gas model with nearest neighbor (NN) interatomic pair interactions. In addition to the NN interactions, cluster interactions are present inside continuous chains of adjasent atoms. The model was initially introduced by Wako and Saitô [1, 2] and by Muñoz and Eaton [3, 4] to understand protein folding. The role of atoms played the peptide bonds. Recently this model was used to describe RNA folding [5]. Furthermore, a similar model was derived in a theory of strained epitaxy [6, 7].

The physical meaning of the cluster interactions is easily understandable in the case of coherent strained epitaxy. Let us assume that besides the attractive NN interaction $v_1 < 0$ the interatomic potential has a rigid core which does not let the atoms approach each other closer than the core diameter d [8]. So if the diameter is larger than the substrate lattice spacing a, the adatoms within an atomic chain will be displaced from the centers of the deposition sites by $u_j \propto f$. The requirement of coherence means that the displacements should be small in order for the displaced atoms remained within the same lattice cell. In general this condition will be violated for sufficiently long chains but in the present study we consider only finite systems and assume that the misfit is sufficiently small for the condition of coherence to be satisfied. In this case the misfit energy of atom j in the harmonic approximation can be estimated as $ku_j^2/2$, where k is the curvature of the substrate potential near its minimum. The atomic displacements within a chain of length l can be found from symmetry considerations as

$$u_{\pm j} = \pm \begin{cases} f(j+1/2) & j = 0, 1, \ldots, l/2-1 \quad l \text{ even} \\ fj & j = 0, 1, \ldots, (l-1)/2 \quad l \text{ odd} \end{cases} . \tag{1}$$

With the use of identities

$$\sum_{j=1}^{m} j = m(m+1)/2$$

and

$$\sum_{j=1}^{m} j^2 = m(m+1)(2m+1)/6$$

the total energy of the chain of length l after some algebra can be calculated as

$$E^{(l)} = V^{(1)} l + v_1(l-1) + (kf^2/24) l(l^2-1), \tag{2}$$

where $V^{(1)}$ is the adsorption energy per atom.

Let us assume that $N_a < N$ adatoms are gathered onto N_a/l equal chains of length l. The total energy of the system in this case will be equal to $(N_a/l)E^{(l)}$. Thus, the energy minimum at fixed N_a will coincide with the minimum of $E^{(l)}/l$. From (2) it is easy to see that such a minimum always exists provided $f \neq 0$. This produces a simple model of self-assembly of size calibrated coherent nanostructures similar to quantum dots.

Formally the hamiltonian (or, more precisely, the configuration dependent free energy) of the WSME model is [1, 2, 3, 4, 6, 7, 9]

$$H^{(N)}_{\mathrm{WSME}} = \sum_{l=1}^{N} \sum_{i=l}^{N} V_i^{(l)} \prod_{k=i-l+1}^{i} n_k, \tag{3}$$

where in the case of epitaxy N is the total number of deposition sites, $n_j = 0, 1$ describes the occupation of site j by the gas atom, $V_i^{(l)}$ are inhomogeneous (i. e., site-dependent) interactions within the continuous atomic chains of length l ending at site i, as can be seen from the product in (3). As was shown in Equation (5) of [10], if chain energies $E^{(l)}$ for all l are known, the values of $V^{(l)}$ in (3) can be found as the discrete second derivative of $E^{(l)}$ with respect to l. Furthermore, to simplify notation we assume the chemical potential μ to be included (with the minus sign) into the parameters $V_i^{(1)}$.

In the case of biopolymers the interpretation of the interactions in the model is different. Firstly, the "atoms" in this case are either the amino-acid residues [1, 2] or the covalent bonds [3, 4, 9] which are assumed to be present in two states: the native and the non-native one corresponding to the values 1 and 0 of a binary variable, respectively [1, 2, 3, 4, 5]. Thus, N can be either the number of peptide bonds connecting $N + 1$ amino-acid residues or the number of the residues themselves. In the more developed bond model $V_i^{(1)}$ is the loss of conformation entropy by the bond in the process of formation of the native state [3, 4, 9]. Cluster interactions $V^{(l)}$ in (3) can be presented in the form [9]

$$V_i^{(l)} = \varepsilon_{ji} \Delta_{ji}, \tag{4}$$

where i and $j = i - l + 1$ are the 1D coordinates of two peptide bonds; $\Delta_{ji} = 1$ if the bonds are in contact with each other and is equal to zero otherwise; ε_{ji} is the inter-residue interaction energy between residues i and $j + 1$. Farther details are given in [9].

The binary matrix Δ_{ji} defines the contact map of a protein in its native state. It depends on the definition of the residue contact. The major parameter here is the cutoff distance between atoms which separates the atoms considered to be in contact from remote atoms. For example, if the distance is chosen to be 8 Å then each residue on average contacts with approximately ten other residues (see Table 1 in [11]). For smaller values of the cutoff chosen in [2] the average number of contacts per residue is $\lesssim 3$ (see Table I in [2]).

The qualitative similarity between the epitaxial and the folding models can be seen on the lattice protein folding model considered in [12]. According to the model, the folding starts at random places in the process of nucleation of a local native structure. The binary bond variables inside the regions are all equal to unity while in other regions the variables are zero. Statistically such behavior can be described by the WSME model.

Despite being 1D, the WSME model has two peculiarities hampering its exact solution. The first is the absence of the translational invariance which makes inapplicable the efficient techniques of the homogeneous case [1, 2, 6, 10]. The other peculiarity is

that the hamiltonian (3) contains long-range interactions so the conventional transfer matrix (TM) method cannot be used. Because of these peculiarities, the exact solution for the WSME model at equilibrium in the inhomogeneous case was found only recently [9]. This solution, on the one hand, greatly facilitates the study of the kinetics of folding [13], on the other hand, it allows for the modeling of the strained epitaxy in inhomogeneous environments, such as alloyed substrates [7]. This latter case is of interest in connection with engineering applications where the 1D nanostructures (such as nanowires, nanomagnets, nanotubes, etc.) may have important applications [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25]. Because in the device environment the wire (for example) may traverse different chemical surroundings, make turns, experience disordered substrate potential due to doping, etc., the interaction parameters describing the model should in general be position-dependent.

In the epitaxial systems, however, the inhomogeneous WSME model describes only 1D chains of atoms or molecules. But in practical applications more than monatomic structures can be needed in order, e. g., to enhance the conductivity of a nanowire or to increase the magnetic moment of a nanomagnet. These structures may consist of several adjacent atomic rows on a terrace of a vicinal surface or on the surface of a nanotube. Such quasi-2D structures can be described with the use of the WSME model only if at least further neighbor pair interactions are added to the hamiltonian (3).

Indeed, let us consider the topology of the deposition sites shown on figure 1. This topology may correspond to the deposition sites on the terrace of a vicinal surface with a rectangular geometry. In this case atoms 2 and 6 or 7 and 11 will be nearest neighbors on the substrate lattice but not along the 1D lattice. But if the substrate has the geometry of triangular lattice with the angle 2–1–5 being equal to 120°, the atoms 1–6 and 7–12 (for example) will constitute additional nearest neighbor pairs. If, farther, the sites in figure 1 are rolled into a cylinder with the chiral vector (4,0), the pairs of sites of type 1 and 4 or 9 and 12 become nearest neighbors too. Other tube topologies will bring together other atoms. An example of the nanotube with chirality (4,1) will be given in section 4. Obviously in all these cases the neglect of the interactions between the atoms on the nearest neighbor sites of the substrate lattice will qualitatively change the physics of the system under consideration. But such interactions in the 1D lattice coordinates will be further neighbor interactions (4-th neighbors in the above case of the tube of rectangular geometry) which do not enter into the WSME hamiltonian (3).

Similar arguments can be applied also to 2- and 3D lattice models of proteins [26, 12]. Muñoz and Eaton noted in this connection [3] that the inclusion of interactions between the bonds belonging to different native stretches into the WSME model not only will improve quantitative description of the kinetics but also "would add considerable flexibility to possible structural mechanisms by producing additional routes between the denatured and native states".

The non-WSME interactions may be of qualitative importance in differentiating between the proteins and the RNA folding. While in some respects being similar, the

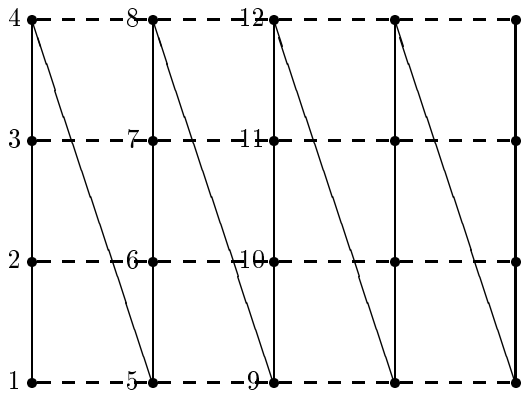

Figure 1. Black points represent the deposition sites on the substrate; solid line with numbers shows a possible mapping of the sites on a 1D lattice with integer coordinates. Dashed lines connect nearest neighbor sites on the substrate lattice with rectangular geometry. These sites are the nearest neighbors on the substrate but the fourth neighbors in the 1D lattice coordinates. For farther information see the text.

folding of the two polymer types differ [27]. In particular, in contrast to proteins, the RNA native state is hierarchical in that the secondary structure is energetically well separated from the tertiary one and can be considered as a collection of base pairings [28, 29, 30]. Because of this, in the standard model of the RNA the energetics is governed by the medium and long range pair interactions [29, 31].

Last but not least, even when the larger distance along the 1D lattice corresponds to larger separation in real space so that further neighbor interactions are small, in some cases they also may cause qualitative phenomena. For example, if the model considered at the beginning of this section is augmented by the substrate mediated repulsive dipole pair interactions [32, 33] or, alternatively, by attractive interaction of some other origin [10] the model, in addition to the self-assembly and size calibration, will simulate the important phenomenon of self-organization of quantum dots into periodic arrays. Thus, there exist many situations when farther neighbor interactions are among the most important ones in the system and cannot be neglected without qualitatively changing the system properties.

But besides the pair interactions, sufficiently strong cluster interactions of non-WSME kind are also usually present in epitaxial systems (see the next section). In this study we present therefore a TM solution of the WSME model augmented by arbitrary short range interactions. Because of the restrictions imposed by the TM technique, in practice the method will be restricted to the interactions of relatively short range. Nonetheless, fairly large radii up to those which in proteomics are classified as the medium- and moderately long-range interactions are feasible to exact treatment within our approach.

In the next section we introduce our extension of the WSME model, in section 3 explain its formal solution and in section 4 present a simple illustrative calculation. In

the concluding section 5 we discuss our results.

2. Extended WSME model

The pair interactions, such as the Coulomb or the van der Waals are the most ubiquitous in nature and their account should be the primary goal in extending the WSME model. But more complex cluster interactions (CIs) can also be important, especially in metallic systems where the pair approximation holds only approximately. For example, *ab initio* calculations show that interactions within atomic trios deposited at the surface may have the same magnitude as the nearest neighbor pair interactions, i. e., to be among the strongest in the system [34, 35]. To account for all possible CIs, in the *ab initio* theories of alloys and epitaxial systems the method of interatomic interaction expansion over a complete set of CIs has been developed [36, 37, 34]. It is pertinent to note that binary alloy is formally equivalent to the lattice gas model. In order for our approach to be compatible with this powerful technique, we developed it for an arbitrary set of CIs restricted only by the maximum values of their interaction radii. This is necessary for the computational tractability of the TM equations.

Let us first consider the most general hamiltonian which includes all possible CIs in the system of size N

$$H_{N-1}^{(N)} = \sum_{C=1}^{2^N-1} W_C n_N^{c_{N-1}} n_{N-1}^{c_{N-2}} \dots n_2^{c_1} n_1^{c_0}, \tag{5}$$

where the subscript $N-1$ denotes the maximum interaction radius. We define the radius of a CI as $r = i_{max} - i_{min}$, where $i_{max(min)}$ is the largest (smallest) index of n_i in the cluster, $c_i = 0, 1$ and, by definition, $n_i^0 \equiv 1$. The CIs in (5) are characterized by sequences of binary digits which can be gathered into the number

$$\bar{C} = (c_{N-1} \dots c_1 c_0)_B, \tag{6}$$

where the bar over a number denotes that its binary representation is meant; the subscript B denotes that the term within parentheses is the binary representation, not the product and W_C is the strength of the corresponding CI. In (5), (3) and below we list the terms in the products in reverse order because in our TM approach it is convenient to number the sites from right to left.

The total number of CIs in hamiltonian (5) is of $\mathrm{O}(2^N)$ which is a huge number for even modest systems of sizes $N \approx 50$ characteristic for the smallest proteins. Only a small part of $\mathrm{O}(N^2)$ of the CIs from (5) enters into the hamiltonian (3). Obviously, in the general inhomogeneous case it would be impossible to take into account all interactions for a system of practical interest. To make the problem manageable, we restrict the extent of the interactions by some maximum radius R.

The extended WSME model we will solve in the next section has the hamiltonian

$$H^{(N)} = H_{\mathrm{WSME}}^{(N)} + H_R^{(N)}, \tag{7}$$

where the second term on the right hand side is defined as the sum of all such terms in (5) that do not contain interactions of radii exceeding R and, besides, in order to avoid double counting these terms should not enter into (3).

3. Recursive transfer matrix solution

In physical terms the main difference between the models represented by hamiltonians (3) and (7) is as follows. In (3) due to the specific form of interactions the energy of any configuration is the sum of energies of the continuous atomic chains (or the stretches of the peptide bonds) it contains. The interactions in (3) become zero as long as atomic clusters are separated by a single empty site. Thus, in the homogeneous case the system can be considered as a mixture of non-interacting molecules of N kinds (different sizes) [10]. Presumably because of this simplicity the homogeneous case was solved much earlier than the general case [1, 2].

The additional term in (7) changes drastically the situation even in the homogeneous case because now not only different chains interact but their interactions are quite nontrivial. For example, in the case of $R = 14$ considered in [35], up to eight islands may be interacting via appropriate CIs. Because of this, there seems to be no way of accounting for all possible situations except through their direct enumeration. In the case of finite range interactions and in 1D this can be done recursively by adding sites to the system one by one.

So let us for the time being neglect in (3) all interactions whose radii exceed R. Because of the finite interaction range, when adding a site to the system consisting of $K \geq R$ sites only the interactions with the last R sites need be taken into account. The accounting can be done with the use of the vector partition function $\vec{Z}^{(K)}$ whose components are the partial traces over all except the last R sites

$$Z^{(K)}_{n_K,n_{K-1},\ldots,n_{K-R+1}} = \mathrm{Tr}_{n_1,n_2,\ldots,n_{K-R}} \exp(-H^{(K)}). \tag{8}$$

Here $H^{(K)}$ is a hamiltonian (7) for a K-site system which contains only interactions within the range not exceeding R. The total partition function is found from (8) as

$$Z^{(K)} = \sum_{\bar{\alpha}=\bar{0}}^{\overline{2^R-1}} Z^{(K)}_{\bar{\alpha}}, \tag{9}$$

where the bar over the number has the same meaning as in (6).

As was shown in Appendix of [35] and will be explained in more detail below, a recurrence relation for $\vec{Z}^{(N)}$ in the number of sites in the system N can be established. This technique is an extension of the methods developed in connection with the 2D Ising model in [38, 39]. Its advantage is that it deals with sparse TMs which provide considerable gain in computational effort in the case of large R.

If we assume that the vector partition function for the system of size $N-1$ is known then the partition function for the size N can be calculated recursively with the use of

the sparse TM as

$$\begin{pmatrix} \circ\circ\ldots\circ\circ \\ \circ\circ\ldots\circ\bullet \\ \vdots \\ \circ\bullet\ldots\bullet\bullet \\ \bullet\circ\ldots\circ\circ \\ \bullet\circ\ldots\circ\bullet \\ \vdots \\ \bullet\bullet\ldots\bullet\bullet \end{pmatrix}^{(N)} = \begin{pmatrix} 1 & 1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 0 & 1 & 1 \\ b_{\bar{0}} & b_{\bar{1}} & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & b_{\bar{2}} & b_{\bar{3}} & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 0 & b_{\overline{2^R-2}} & b_{\overline{2^R-1}} \end{pmatrix}_N \begin{pmatrix} \circ\circ\ldots\circ\circ \\ \circ\circ\ldots\circ\bullet \\ \vdots \\ \circ\bullet\ldots\bullet\bullet \\ \bullet\circ\ldots\circ\circ \\ \bullet\circ\ldots\circ\bullet \\ \vdots \\ \bullet\bullet\ldots\bullet\bullet \end{pmatrix}^{(N-1)}, \quad (10)$$

where the column vectors correspond to $\vec{Z}^{(N)[(N-1)]}$, the empty and filled circles describe the empty ($n_i = 0$) or filled ($n_i = 1$) sites in the subscripts of the partial partition functions in (8) and the subscript N of the TM is the site index for all $b_{\bar{\alpha}}$ entering the matrix. We note that we use the same symbol N for the system size and for the recurrent relation to stress that at every iteration we obtain the (vector) partition function of a system corresponding to some size N.

In the case of finite-range interactions the structure of TM in (10) is easily understood. Having added site N to the system consisting of $N-1$ sites we first have to account for the interaction of this site with the rest of the system and then take the trace over the $(N-R)$-th site because with the radius of interactions being R all interactions of this site with the rest of the system have already been taken into account. Taking the trace amounts to adding with appropriate weights two $Z^{(N-1)}$ differing by the filling of site $N-R$. In the case of the empty site N the weights are equal to unity because the empty site does not interact with anything. These terms occupy the upper half of the TM (10). The lower half of the matrix contains the terms corresponding to the interaction of the *occupied* site N with the rest of the system. The term

$$b_{\bar{\alpha}N} = \exp(-\Delta E_{\bar{\alpha}N}/k_B T) \quad (11)$$

is the Boltzmann weight corresponding to the interaction of the atom at site N with the configuration of atoms corresponding to $Z_{\bar{\alpha}}^{(N-1)}$; $\Delta E_{\bar{\alpha}N}$ in (11) is the energy of interaction of the atom with configuration $\bar{\alpha}$.

Now, what have to be changed in order to include the arbitrary range interactions of the WSME type into the recursion scheme (10)? It turns out that only the last equation need be modified. This is because the hamiltonian (3) contains only the interactions inside continuous chains. But all components of the state vector $\vec{Z}^{(N-1)}$ except the last one contain at least one empty site among the last R sites. Therefore, the extent of the chain interactions is restricted by the distance to the nearest empty site and thus is smaller than R.

In the last component, however, all sites are filled. So when adding an additional N-th site filled with an atom we do not know which interactions of the WSME type should be taken into account as the last R atom may belong to a chain of any length—from R to $N-1$. We overcome this difficulty in a straightforward manner by simply

taking into account all the possibilities. Namely, we replace the last two-term equation in the set (10) with the sum over all configurations where the last R sites belong to a chain of length greater or equal to R. This can be achieved with the use of the component $Z^{(M)}_{2^R-2} \equiv \bullet\bullet\ldots\bullet\bullet\circ$ with all sites except the first one being filled. Note that the positions are counted from right to left. When adding chains of different lengths to this component we can control the total chain length and thus know which interactions from (3) should be taken into account.

Formally this is done as follows. In the course of the recursive solution we keep the array of components $Z^{(M)}_{2^R-2}$ for $M = R-1, \ldots, N-1$ (the explanation of the term $R-1$ is given below); in another array we gather the chain energies $E^{(l)}_N$. These account for all interatomic interactions entering (7) *inside* the chains of length l ending at site N. The chains are assumed to be isolated so no interchain interactions enter $E^{(l)}_N$. By attaching a chain of length $N - M + R - 1$ to the configuration corresponding to $Z^{(M)}_{2^R-2}$ which amounts to multiplying the latter by the corresponding Boltzmann factor, we obtain a configuration with a continuous chain of atoms starting on site $M - R + 1$ and ending on site N. As is seen, the $(R-1)$-atom chain in $Z^{(M)}_{2^R-2}$ ending at site M and the chain ending at N overlap at sites inside the (sub)chain of length $R-1$. In the equation below this double counting is taken care of by the division by the necessary Boltzmann factor corresponding to the chain of length $(R-1)$:

$$Z^{(N)}_{2^R-1} = \sum_{M=R-1}^{N-1} \exp[-E^{(N-M+R-1)}_N/k_BT]\tilde{Z}^{(M)}_{2^R-2}, \tag{12}$$

where

$$\tilde{Z}^{(M)}_{2^R-2} = Z^{(M)}_{2^R-2}/\exp[-E^{(R-1)}_M/k_BT]. \tag{13}$$

The meaning of (12) is simple: the component of $\vec{Z}^{(N)}$ with the last R sites being filled is obtained as the sum of all possible configurations having the chains of length $R \le l \le N$ as their end sites. As is easy to see, the factor $\tilde{Z}^{(M)}_{2^R-2}$ is sufficient for accounting for all short-range interactions of the chain of length $N-M+R-1$ with the rest of the system because the atoms at sites $M+1$ and larger cannot reach the atoms beyond $M-R$ due to the finite interaction range. The only remaining problem is connected with the longest chain of length N which should comprise the whole system because $Z^{(R)}_{2^R-2}$ starts with an empty site. This difficulty is overcome by initializing the recurrence (10) with $\vec{Z}^{(R-1)}$, i. e., with the system containing only $R-1$ sites instead of R. The fillings of these sites correspond to the last $R-1$ sites in the vectors in (10), i. e., the rightmost column in these vectors should be crossed out so the component $Z^{(R-1)}_{2^R-2}$ has all its sites filled. The components corresponding to the empty crossed out sites are calculated as the conventional Boltzmann factors while in the cases when the omitted site was filled they are all set to zero. The validity of this initialization of the recurrence (10) can be proven either by a straightforward calculation of $\vec{Z}^{(R)}$ via one iteration step and comparing it with $\vec{Z}^{(R)}$ calculated straightforwardly, or by associating with the crossed out column a fictitious 0-th site which has an infinite on-site energy. Thus, on the one

hand, the initial $\vec{Z}^{(R-1)}$ corresponds to the system of size R $(0,\ldots,R-1)$; on the other hand, the components corresponding to this site being filled are all zero, as suggested above.

4. Illustrative calculation

As can be seen from (12), formally our algorithm is quadratic in the system size N. This means that for sufficiently large systems the calculations may become prohibitively difficult to perform. The sizes of biopolymers met in nature, however, are restricted [40]. Because from practical point of view of major interest are natural biological molecules, we will restrict our discussion to this case. A typical protein consists from about 500 amino acid residues [31]. So in order to assess numerical performance of our algorithm we consider for simplicity an epitaxial model of this size. The epitaxial systems of similar sizes are of interest also for the nanoengineering. Because the devices of sizes in tens of nanometers (hundreds of atoms) are efficiently modeled in the framework of continuum approximations [41, 42, 43], our approach may be useful in studies of smaller few-nanometer structures [14]. Thus, the length in 500 atomic diameters (about 100 nm) is, presumably, an upper limit of interest for the atomic simulations of epitaxial systems (the issue of their width will be discussed below).

We consider coherent strained epitaxy on the surface of a finite size screw (4,1) nanotube with rectangular substrate lattice geometry (see figure 1) with homogeneous interactions and consisting of 500 deposition sites. This geometry was chosen because the diameter four would correspond, *inter alia*, to a model of the α-helix similar to that considered in [44] but with additional pair interactions. This may be used to model the helices with different interbond interactions to better describe their properties. The (4,1) topology means that sites 2 and 6 or 7 and 11 in figure 1 are nearest neighbors along the direction parallel to the tube axis while the sites along the solid line are all equivalent. For example, the interaction between atoms at sites 2 and 3 is the same as between those at sites 8 and 9 because all points along the line belong to a helix. The potential of the substrate (the tube surface) is periodically corrugated along the helix and will be treated in the harmonic approximation (2), as discussed in the Introduction.

Besides the positive misfit energy, the atoms in our model experience, apart from the NN interaction v_1, small attraction between the first and the second neighbors along the helix. Because of the homogeneity of the model, the nomenclature of (5) is superfluous, so below we denote this interaction as v_2. Besides, v_4 will designate the repulsion between the atoms which are the fourth neighbors along the helix but are NN on the substrate surface (see figure 1). This model qualitatively describes the large misfit systems studied in [25] and [24]. In these papers it was found that while on the tubes of large diameters the interaction between the nearest neighbor adatoms is repulsive along both directions, on those of small diameters the interaction along the high curvature direction became attractive due to the increased interatomic distance. But the interaction in the direction of small curvature along the tube axis remains

repulsive even for the small-diameter tubes.

In the explicit calculations below we used the following values in units of the NN attractive interaction $v_1 < 0$:

$$v_2 = 0.3v_1, \quad v_4 = 0.2|v_1| \quad \text{and} \quad kf^2 \approx 6.8 \cdot 10^{-3}|v_1|. \tag{14}$$

The energy of an isolated chain of length l can be obtained from (2) by adding to it the terms due to v_2 and v_4

$$E^{(l)} = \sum_{i=1,2,4} (l-i)_{>0} v_i + (kf^2/24) l(l^2-1) - \mu l, \tag{15}$$

where the subscript > 0 means that only positive values of $(l-i)$ contribute and $V^{(1)}$ in (2) was set equal to $-\mu$.

Numerical values of the pair interactions (14) were chosen in such a way that in the absence of misfit ($f = 0$) the reduced energy $E^{(l)}/l$ did not have a global minimum at finite value of l so that the system were of phase separation type. This means that the atoms at low temperature tend to gather into one cluster. In the presence of the misfit, however, $E^{(l)}/l$ has a local minimum at $l = 12$. This choice means that the chain makes three turns around the tube which approximately corresponds to the structure of the typical α-helix. This model was solved with the recursive technique of the previous section at three different temperatures for the system consisting of 500 sites at half coverage (250 atoms). In figure 2 are shown the size distributions of chains of different lengths on the surface of a cylinder under consideration. As is seen, at the highest temperature the size distribution is similar to the random distribution of atoms while at the lowest temperature it exhibits very good size calibration with $\gtrsim 96\%$ of atoms belonging to chains of lengths 11, 12 or 13. Thus, in the presence of the misfit the atoms gather into chains of about 12 atoms each. This result is in accord with the theory [45] but it was not obvious from the start because the interisland interactions are known to shift the calibrated size from its noninteracting value at the minimum of $E^{(l)}/l$.

All calculations were performed on a modern Intel® processor with the use of Python scripts. This choice was motivated by the problem of numerical underflow or overflow which appeared in the calculations. The problem is rooted in the exponential scaling of the partition function with the system size: $Z^{(N)} \sim \exp(N\phi)$, where ϕ is the reduced (per site) grand potential. Because of this, at sufficiently large N $Z^{(N)}$ may acquire arbitrarily large or small values which exhaust any fixed numerical range. In Python libraries, however, there exists the module `decimal` which allows for calculations with extremely small and/or extremely large numbers.

When present, this problem severely hampers the calculations. For example, in the above model with $N = 500$ and the parameters shown in figure 2 calculation of $Z^{(500)}$ required only a fraction of a second because the conventional double precision arithmetic was sufficient: The maximum values of $Z^{(500)}$ were of $\mathrm{O}(10^{101})$. But this means that the system with, e. g., 5000 sites will have $Z^{(5000)} \sim \mathrm{O}(1010)$ which in our approach would require the use of the module `decimal`. Indeed, explicit calculation gave $Z^{(5000)} \approx 1.5 \cdot 10^{1011}$. This calculation took almost 35 minutes which is about

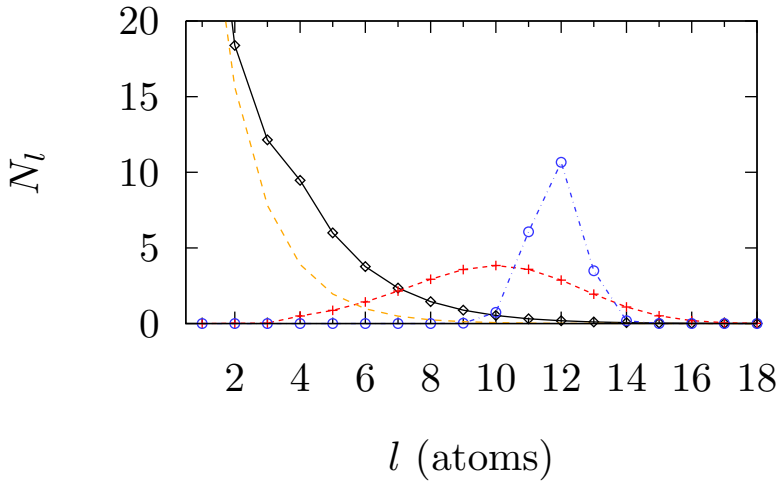

Figure 2. Low temperature size calibration of self-assembled atomic clusters in the model described in the text. The number of clusters of different sizes at different temperatures: ◇—$T = |v_1|$, +—$T = 0.1|v_1|$ and ○—$T = 0.01|v_1|$. The curves are guides to the eye except the monotonous dashed curve which describes random coverage.

150 times longer than the calculation at this size without the module. Repeated with the parameters where the double precision was sufficient it took only about 14 seconds. Thus, the software realization of high-precision arithmetics costs more than two orders of magnitude in performance. If this is characteristic for all such software, much better choice is to use the quadruple precision realized in some C/C++ and Fortran compilers. In addition to much smaller overhead due to higher precision, the compiled languages offer additional speed up in about two orders of magnitude in comparison with the interpreted languages such as Python. In this way the calculations with large biopolymers should be very fast. For example, the calculation of $Z^{(35000)}$ with the Python script in the case when the double precision was sufficient took less than 12 minutes which with the compiled language should take only a few seconds.

It should be reminded that all calculations were performed for $R = 4$. As can be seen from (10) and (12), the equations have the form of scalar products of vectors of sizes 2^R and N, respectively. This means that the calculations are trivially parallelizable, so the execution speed at large R will depend on the number of processors available for the calculation and on the performance of one processor. The latter can be assessed from the model calculation of $Z^{(500)}$ with $R = 20$ or $\mathrm{O}(10^6)$ equations in the set (10) with the use of a Python script which lasted about 15 minutes. This means that with a compiled language the time of the calculation will measure in seconds. We estimate that the calculations with R in the range $\lesssim 40$ should be feasible on modern supercomputers.

5. Discussion

In this paper we presented a transfer matrix solution of the WSME model extended to account for arbitrary short range interactions. The transfer matrix approach is, in principle, a universal technique capable of solving any lattice problem with short range interactions. In practice, however, it is restricted to relatively small interaction radii due to the exponential growth of the computational effort with R. This restriction does not allow the method to be considered as a universal tool for obtaining exact solutions in dimensions D>1 because in statistical mechanics one is usually interested in the thermodynamic limit which corresponds to $R \to \infty$ and thus is unaccessible to the TM technique in truly 2- or 3D systems [38, 39, 35].

The natural biopolymers, however, though sometimes very large, are restricted in their maximum size, so their characteristic dimensions are also finite. According to current nomenclature the radii of short and medium range interactions in proteins do not exceed 20 residues [11]. As we saw, this case causes no difficulty even for a single processor computer. On a supercomputer with tens to hundreds parallel processors even the moderately long-range interactions of the extent $R \lesssim 40$ studied in [12] should cause no problems. Thus, in the case of proteins our approach potentially allows for the *exact* solution of the extended WSME model with arbitrary medium- and moderately long-range interactions. According to [11] (see their fig. 4), the interactions with ranges exceeding 40 constitute only about 5% of all interactions. Thus, the technique developed in the present paper allows to improve up to 95% of all interactions.

The RNA molecules are less amenable to the study within our TM technique because the double-stranded nature [27, 31] of the polymer makes the pair interactions very long-ranged, up to the total molecule length when the first and the N-th nucleotides pair. Therefore, the pair interactions in the ranges extending not farther than about 20 pairs along the stem away from the hairpin loops can be treated within our approach. The pseudo knots formed by nearby hairpin heads are other potential candidates for the description with non-WSME interactions [29]. An alternative way of describing the nucleotide interactions is to include them into the WSME part [5].

In the case of epitaxial systems the maximum interaction radius $R \lesssim 40$ lattice units restricts the application of the method to the nanotubes of similar and even lesser circumference [35]. In this connection it is pertinent to note that $R \approx 20$ approximately corresponds to the upper limit of the tube size when the high curvature of small diameter nanotubes can qualitatively change the ordering of adsorbates consisting from large atoms [25]. In the case of deposition on the terraces the upper limit of width ($\lesssim 40$ atomic rows) even exceeds the width of the terraces (16 rows) used in [21] for epitaxial growth of magnetic nanostructures. Such a restriction of the accessible widths is not very serious from the practical point of view. In the case of wide nanowires in tens of atoms the atomic resolution is not very important because an error in a few atoms can be neglected in most cases. Nanostructures of such sizes can be efficiently simulated within continuum approximations [41, 42, 43].

Furthermore, the short range interactions can be used to describe the substrate propagated elastic dipole-dipole interaction in 1D model of strained epitaxy proposed in [6]. The dipole-dipole interaction behaves as the inverse cube of the distance [32, 33] and so at $R \sim 10 - 20$ in 1D systems can be neglected in most cases.

It should be mentioned that the direct push interaction [8] leading to the interactions of the WSME type in 1D or on the screw tubes is not operative in wires on the terraces of width greater than two (on the non-rectangular substrate the wires consisting of two rows may still contain such a contribution). In this case the origin of some of the WSME type interactions can be different. For example, the interaction corresponding to the largest cluster containing all atoms differentiates two cases: the fully filled terrace and the terrace with one vacancy. Such an interaction may account for the volume contribution to the vacancy formation enthalpy. Thus, there is enough interesting epitaxial systems (and we mentioned only a few of them) which can be simulated with the exact transfer matrix technique developed in the present paper.

Acknowledgments

The authors acknowledge CNRS for support of their collaboration. One of the authors (V.I.T.) expresses his gratitude to Université de Strasbourg and IPCMS for their hospitality. We thank M. Alouani for a critical reading of the manuscript.

References

[1] Wako H and Saitô N 1978 *J. Phys. Soc. Jpn.* **44** 1931
[2] Wako H and Saitô N 1978 *J. Phys. Soc. Jpn.* **44** 1939
[3] Muñoz V and Eaton W A 1999 *Proc. Natl. Acad. Sci.* **96** 11311
[4] Muñoz V, Henry E R, Hofrichter J and Eaton W A 1998 *Proc. Natl. Acad. Sci.* **95** 5872
[5] Imparato A, Pelizzola A and Zamparo M 2009 *Phys. Rev. Lett.* **103** 188102
[6] Tokar V I and Dreyssé H 2003 *Phys. Rev. B* **68** 195419
[7] Tokar V I and Dreyssé H 2004 *J. Phys.: Condens. Matter* **16** S2203
[8] Tokar V I and Dreyssé H 2008 *Phil. Mag.* **88** 2747
[9] Bruscolini P and Pelizzola A 2002 *Phys. Rev. Lett.* **88** 258101
[10] Tokar V I and Dreyssé H 2003 *Phys. Rev. E* **68** 011601
[11] Gromiha M M and Selvaraj S 2004 *Prog. Biophys. Mol. Biol.* **86** 235
[12] Abe H and Wako H 2009 *Physica A* **388** 3442
[13] Zamparo M and Pelizzola A 2006 *Phys. Rev. Lett.* **97** 068106
[14] Barth J V, Costantini G and Kern K 2005 *Nature* **437** 671
[15] Bowler D R 2004 *J. Phys.: Condenc. Matter.* **16** R721
[16] Owen J H G, Miki K and Bowler D R 2006 *J. Mater. Sci.* **41** 4568
[17] Gonzàlez C, Snijders P C, Ortega J, Pérez R, Flores F, Rogge S and Weitering H H 2004 *Phys. Rev. Lett.* **93** 126106
[18] Himpsel F J, Ortega J E, Mankey G J and Willis R F 1998 *Adv. Phys* **47** 511
[19] Gambardella P, Brune H, Kern K and Marchenko V I 2006 *Phys. Rev. B* **73** 245425
[20] Negulyaev N N, Stepanyuk V S, Hergert W, Bruno P and Kirschner J 2008 *Phys. Rev. B* **77** 085430
[21] Repain V, Baudot G, Ellmer H and Rousset S 2002 *Europhys. Lett.* **58** 730
[22] Zhou C, Kong J, Yenilmez E and Dai H 2000 *Science* **290** 1552

[23] Wang Z, Wei J, Morse P, Dash J G, Vilches O E and Cobden D H 2010 *Science* **327** 552
[24] Yang X and Ni J 2004 *Phys. Rev. B* **69** 125419
[25] Lueking A D and Cole M W 2007 *Phys. Rev. B* **75** 095425
[26] Salvi G and P De Los Rios 2003 *Phys. Rev. Lett.* **91** 258102
[27] Thirumalai D and Hyeon C 2005 *Biochemistry* **44** 4957
[28] Schuster P, Fontana W, Stadler P F and Hofacker I L 1994 *Proc. R. Soc. Lond. B* **255** 279
[29] Orland H and Zee A 2002 *Nucl. Phys. B* **620**[**FS**] 456
[30] Wolfsheimer S, Burghardt B, Mann A and Hartmann A K 2008 *J. Stat. Mech.* P03005
[31] Bundschuh R and Hwa T 1999 *Phys. Rev. Lett.* **83** 1479
[32] Lau K H and Kohn W 1977 *Surf. Sci.* **65** 607
[33] Marchenko V I and Parshin A Y 1980 *Sov. Phys. JETP* **52** 129
[34] Luo W and Fichthorn K A 2005 *Phys. Rev. B* **72** 115433
[35] Tokar V I and Dreyssé H 2009 *Preprint* arXiv:0912.2680v2
[36] Sanchez J M, Ducastelle F and Gratias D 1984 *Physica* **128A** 334
[37] Ducastelle F 1991 *Order and Phase Stability in Alloys.* North-Holland, Amsterdam
[38] Kramers H A and Wannier G H 1941 *Phys. Rev.* **60** 252
[39] Domb C 1949 *Proc. Royal Soc. Series A* **196** 36
[40] URL `http://www.pdb.org`
[41] Tu Y and Tersoff J 2007 *Phys. Rev. Lett.* **98** 096103
[42] Grima R, DeGraffenreid J and Venables J A 2007 *Phys. Rev. B* **76** 233405
[43] Eggleston J J and Voorhees P W 2002 *Appl. Phys. Lett.* **80** 306
[44] Flammini A, Banavar J R and Maritan A 2002 *Europhys. Lett.* **58** 623
[45] Priester C and Lannoo M 1995 *Phys. Rev. Lett.* **75** 93